p–ISSN: 2723 - 6609 e-ISSN: 2745-5254
Vol. 5, No. 7 July 2024 http://jist.publikasiindonesia.id/

Comparison of Linear Regression and Random Forest Algorithms for Premium Rice Price Prediction (Case Study: West Java)

Irfan Rasyid Muchtar[1*], Afiyati[2]
Universitas Mercubuana Jakarta, Indonesia
Email: ir16151@gmail.com[1*], afiyati.reno@mercubuana.ac.id[2]

*Correspondence

ABSTRACT

Keywords: Premium Rice Prices, Linear Regression, Random Forest.

The staple food commodity that is crucial to the Indonesian society is rice. Rice often experiences fluctuations in prices. These fluctuations can be predicted using machine learning methods. This research aims to evaluate the accuracy of machine learning algorithms in predicting premium rice prices in the West Java Province, Indonesia. Two methods used in this study are Linear Regression and Random Forest. The dataset used consists of 6096 records from the Indonesian Food Commodity Management Agency. The evaluation results show that the Random Forest algorithm has an accuracy rate of 98.69%, while the Linear Regression algorithm has an accuracy rate of 95.08%. Based on these results, it is concluded that the Random Forest algorithm is more effective in predicting premium rice prices in the West Java Province compared to the Linear Regression algorithm.

Introduction

Rice is one of the staple food commodities that is very popular among the people of Indonesia. Most of the Indonesian population consumes rice as a daily staple food (Hasibuan, Febjislami, and Suliansyah 2022) Therefore, the availability of sufficient rice supply is the main task of the government to maintain national food security. However, rice prices often fluctuate and are influenced by various factors, such as production levels, demand, availability, weather, government policies, and other factors. (Hanim 2016) These price fluctuations can hurt the agricultural and consumer sectors, as well as affect economic stability and national food security.(Ruvananda and Taufiq 2022)

Given the importance of rice in the daily lives of Indonesian people, rice price prediction with high accuracy is very important. Accurate predictions can help the government in taking appropriate policies to maintain price stability and rice availability, as well as protect farmers and consumers from the negative impact of price fluctuations.

The purpose of this study is to compare the performance of the Linear Regression and Random Forest algorithms in predicting premium rice prices in West Java Province. Specifically, this study aims to:

1. Determine the accuracy level of the Linear Regression model in predicting premium rice prices in West Java
2. Determine the level of accuracy of the Random Forest model in predicting the price of premium rice in West Java.
3. Analyze the advantages and disadvantages of each algorithm in the context of premium rice price prediction.
4. Identify the factors that most significantly affect the price of premium rice based on the results of the model.
5. Providing recommendations on more effective and efficient prediction models to be applied in practical contexts in West Java.

Research Methods

At this stage, the steps in the process of implementing the Random Forest method will be compared with the Linear Regression method. The purpose of this study is to find out the performance results of the Random Forest and Linear Regression methods in predicting premium rice prices.

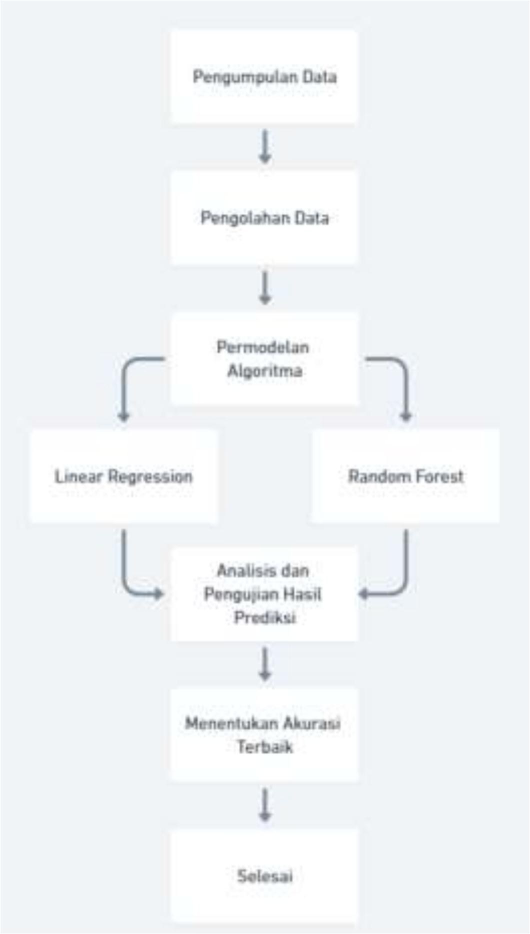


Figure 1 Research Stages

Data Collection

The first step in this study is data collection to predict the price of premium rice. The data that will be used is premium rice price data from the Indonesian Food Management Agency, specifically from the province of West Java. The dataset consists of 6096 entries with various variable components including Date, Rice Harvest Area, Rice Price, GKP (Harvested Dry Rice) Price at the Farmer Level, GKP (Harvested Dry Rice) Price at the Milling Level, and GKG (Dry Grain Flour) Price at the Milling Level. This comprehensive dataset will provide a solid foundation for building and validating the predictive models. By analyzing these variables, we aim to identify significant patterns and relationships that can accurately forecast the price of premium rice, ultimately contributing to better decision-making processes for stakeholders involved in rice production and distribution. The details of this dataset can be seen in Table 1.

Table 1 Dataset

	DATE	RICE HARVEST AREA	GKPTP	GKPTPG	GKGTP	RICE
0	6/1/2022	10	4140	4410	5230	9970
1	6/2/2022	10	4160	4420	5250	9970
2	6/3/2022	10	4180	4410	5230	9890
3	6/4/2022	10	4160	4410	5230	9760
4	6/5/2022	10	4180	4420	5230	9840
1011	5/15/2024	10	5430	5770	6710	13260
1012	5/16/2024	10	5420	5720	6740	13290
1013	5/17/2024	10	5450	5720	6740	13370
1014	5/18/2024	10	5480	5760	6810	13400
1015	5/19/2024	10	5540	5780	6870	13520

Data Processing

The next stage is data processing or preprocessing of the data to be used. At this stage, the data is cleaned or data is cleaned to clean the data that contains noise. Furthermore, the data was processed to determine the correlation between features using heat maps.

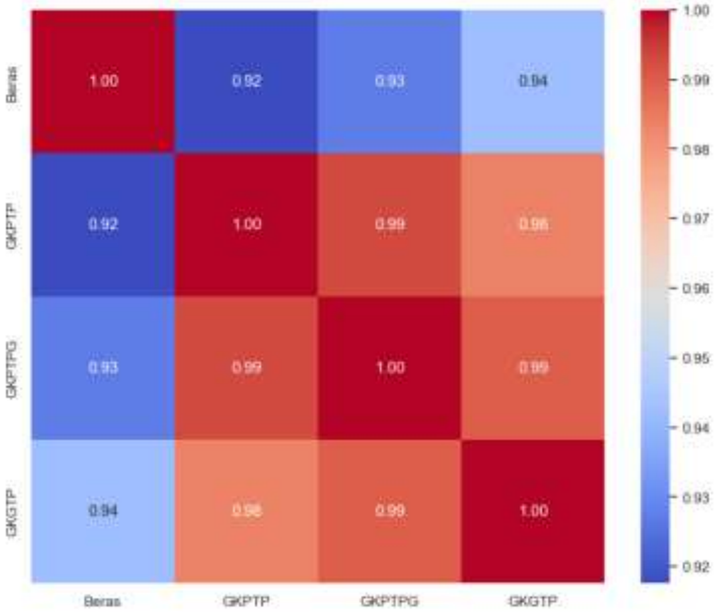


Figure 2 Heatmap

Algorithmic Modeling

Linear Regression

Linear Regression is a method for predicting the value of a variable based on the value of another variable, assuming the relationship is linear. This method uses mathematical equations to describe the relationship. The regression coefficient in the equation shows the magnitude of the influence of the independent variable on the dependent variable. (Supriyanto, Ilhamsyah, and Enri 2022)

$$Y = \alpha + \beta 1X1 + \beta 2X2 + ... + \beta kXk + \varepsilon$$

Y: Dependent variable (the value to be predicted)
X: Independent variable (the value used to predict Y)
α: Constant (Y value when X = 0)
β: Regression coefficient (slope of the regression line)
ε: Error (the difference between the actual value Y and the predicted value Y)

Steps of the Linear Regression Method

a. Data Collection

Collect the necessary data for dependent and independent variables.

b. Data Preparation

Separate the data into two sets: the training set and the testing set. Normalize or standardize data if needed.

c. Model Formation

Use the training data to build a Linear Regression model. Calculate the regression coefficient (β) using the least squares method.

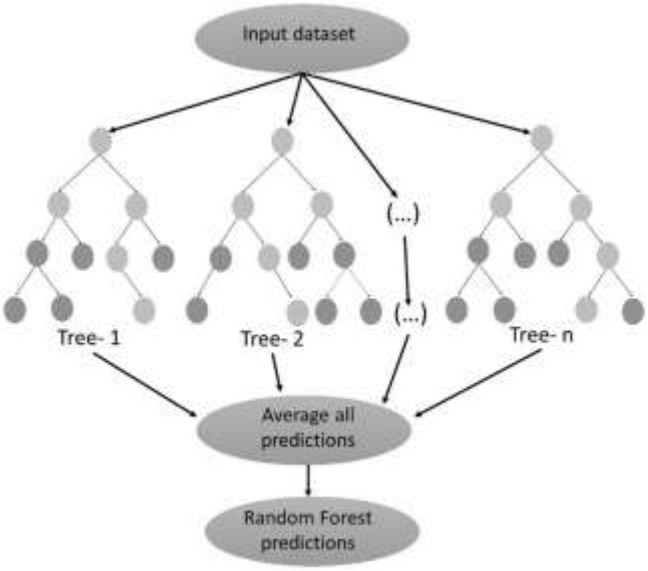
d. Model Validation

Use test data to test the accuracy of the model. Calculate the error value (ε) as the difference between the predicted Y value and the actual Y value. Predictions. Use the regression model that has been built to predict the Y value based on the X value.

Random Forest

Random Forest is a machine learning algorithm used to predict outcomes based on a set of rules created based on the data provided. This algorithm combines the results of multiple Decision Trees to achieve a single, more accurate result. Random Forest is commonly used to solve problems related to classification, regression, and so on. (Saadah and Salsabila 2021).

The Random Forest algorithm is a machine learning algorithm that combines the results (outputs) of multiple Decision Trees to achieve a single, more accurate result. Decision Tree is a non-parametric machine learning algorithm that is shaped like a tree structure. The Decision Tree begins with one node (root node) representing the initial question or decision, the next node (internal node/decision node) represents the answer to the previous question, the branch representing the decision path choice, and the leaf (leaf node/terminal node) representing the final decision. (Karami et al. 2021)



Gambar 3 Diagram Random Forest

The algorithm of Random Forest works in two phases, namely:

1. Building a Decision Tree

In this phase, the Random Forest algorithm will build several Decision Trees. Each Decision Tree will be built from a different subset of training data. (Bsoul, Qusef, and Abu-Soud 2022) On each subset of data, several features will be randomly selected to be used in building the Decision Tree. This is done to reduce the dependency between Decision Trees so that each Decision Tree will learn from different data. The formula used in this phase is

$$h(x) = argmax\_c\ P(c|x, T)$$

Where:

h(x) is the decision function of the *decision tree*

c is the class of data *x*
P(c|x, T) is the probability of *class c* for data *x* based on *decision tree T.*
In this phase, the *Random Forest* algorithm will perform the following steps:

Select a subset of training data

The Random Forest algorithm will randomly select a subset of the training data from the original training data. The number of selected subsets of data will be determined by the estimator parameter. (Shahini and Grgurić 2021)

Select a feature

For each subset of data, the Random Forest algorithm will randomly select several features to use in building the Decision Tree. The number of selected features will be determined by max_features parameters.(Sedaghat et al. 2022)

Buat Decision Tree

The Random Forest algorithm will build a Decision Tree from a selected subset of data and features.

Making predictions

In this phase, the Random Forest algorithm will make predictions for the new data based on the results of several Decision Trees. The final prediction will be determined by counting the most votes from the prediction results from each Decision Tree. This phase will use the following formula:

$$y = argmax\_c \, P(c|x, T\_1), P(c|x, T\_2), ..., P(c|x, T\_n)$$

Where:
y is the class prediction of the x data
c is the class of data *x*
P(c|x, T_i) is the probability of *class c* for data *x* based on *the decision tree T_i.*

Analysis and Testing of Prediction Results

To find out the accuracy level of the prediction system used, we can see from the error level, namely how far the difference between the predicted data and the actual data is. (Putra et al. 2022)

There are several ways to measure accuracy in a prediction system including:

Coefficient of determination (R2)

The R2 value shows how strong the relationship between the independent variable (predicted factor) and the dependent variable (predicted result) is. The R2 value ranges between and 1, with a value of 1 indicating the strongest relationship. On:: $= (\mathrm{yi} - \hat{\mathrm{y}_\mathrm{i}})^2 \,/\, \Sigma(\mathrm{yi} - \bar{\mathrm{y}})^2)$

...yi adalah nilai data aktual
yî is the prediction value
ȳ is the average of the actual data values

Root Mean Square Error (RMSE)

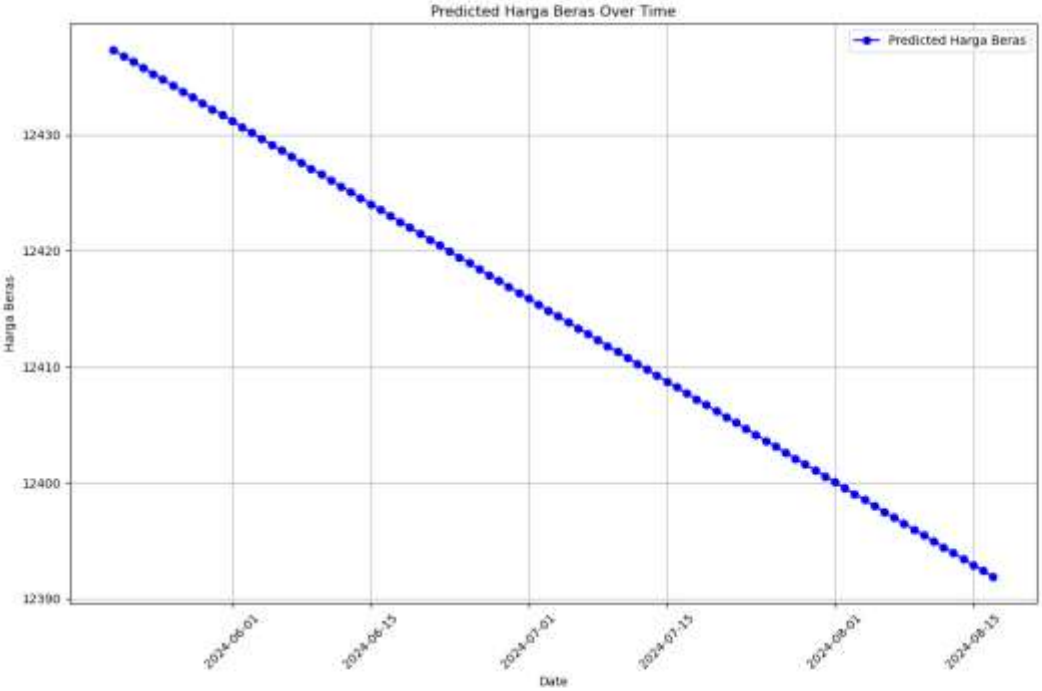
The RMSE value shows how big the average prediction error is. A small RMSE value indicates higher prediction accuracy.

To calculate the RMSE value we can use te formula: $\Sigma(\mathrm{y})^2) \,/\, \mathrm{n})$
yi is the actual data value
yî is the prediction value

n is the amount of data

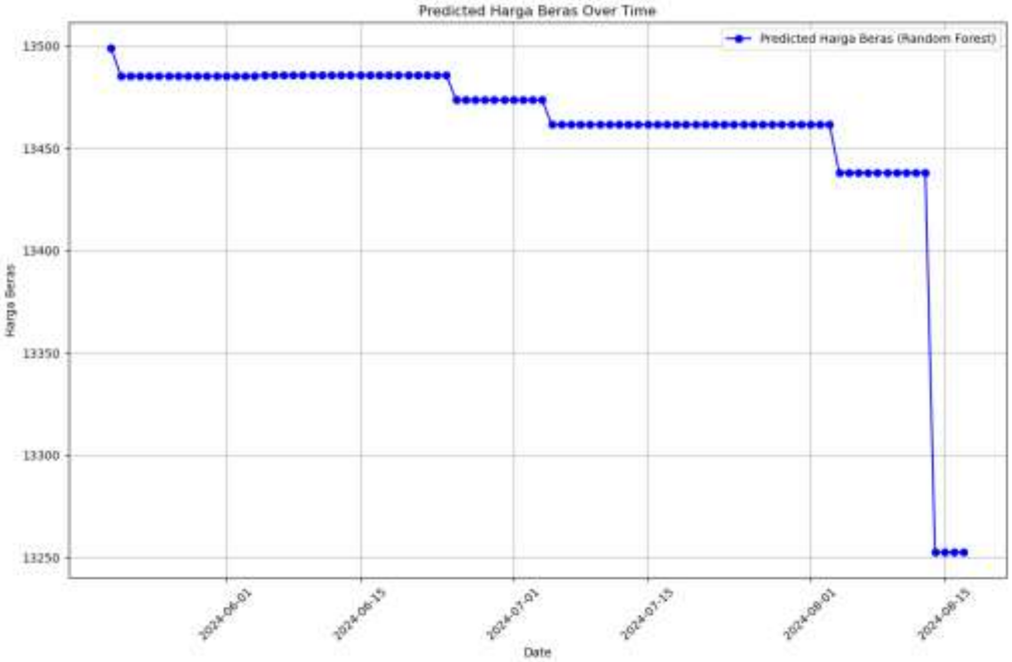
Results and Discussion

Linear Regression Prediction Results



The prediction results from the Linear Regression algorithm show that there will be a decrease in rice prices in the next 3 months in 2024. This can be seen from the visualization of the prediction results of the Linear Regression algorithm which indicates a downward trend in prices. This prediction provides an initial overview of possible changes in rice prices that stakeholders need to anticipate.

Random Forest Prediction Results



Hasil prediksi dari algoritma Random Forest juga menunjukkan bahwa akan terjadi penurunan harga beras pada tahun 2024. Algoritma ini, yang lebih kompleks dan mampu menangkap hubungan non-linier antara variabel, memberikan konfirmasi tambahan terhadap prediksi yang dihasilkan oleh Linear Regression.

Model Evaluation

Algorithm	R2	MAP	RMSE	MAE
Random Forest	0.97	81651	285	160
Linear Regression	0.89	295973	544	397

R2 (R-Squared):

The R2 score of the Random Forest algorithm is 0.97, which indicates that the model can account for 97% of the variation in the data. In contrast, Linear Regression was only able to explain 89% of the variation in the data with an R2 score of 0.89. This indicates that Random Forest has better predictive abilities than Linear Regression. (Estiasih, Waziroh, and Fibrianto 2016)

MSE (Mean Squared Error):

The MSE value of Random Forest is 81651, much lower compared to Linear Regression which has an MSE value of 295973. A lower MSE indicates that the prediction of Random Forest is closer to the actual value compared to Linear Regression. (Lailatul Nikmah et al. 2022)

RMSE (Root Mean Squared Error):

The RMSE value of Random Forest is 285, lower than the RMSE Linear Regression of 544. This confirms that the predictions from Random Forest are more accurate. (Hasanah, Farida, and Yoga 2022)

MAE (Mean Absolute Error):

The MAE value of Random Forest is 160, which is lower compared to the MAE Linear Regression of 397. A lower MAE indicates that Random Forest produces predictions with a smaller absolute error mean.

The evaluation results of the two models show that the R2 Score of the Random Forest algorithm is higher at 0.97 compared to the Linear Regression which only gets an R2 score of 0.89, the MSE, MAE, and RMSE of the Random Forest algorithm are also smaller compared to the Linear Regression, this shows that the Random Forest model provides predictions that are closer to the actual value.

Conclusion

Based on the results of the predictions of the two algorithms, it can be concluded that the price of premium rice will decrease in the next 3 months in 2024. In addition, of the two algorithm models used, the Random Forest algorithm is superior in predicting the price of premium rice. The Random Forest algorithm gets an accuracy score of 97%. This shows that the algorithm has better performance compared to the Linear Regression algorithm which only gets an accuracy score of 89%.

Bibliography

Bsoul, Mohammad A., Abdallah Qusef, and Saleh Abu-Soud. 2022. “Building an Optimal Dataset for Arabic Fake News Detection.” *Procedia Computer Science* 201(C):665–72. doi: 10.1016/j.procs.2022.03.088.

Estiasih, T. .. Waziroh, and K. Fibrianto. 2016. *Kimia Dan Fisik Pangan Bumi Askara. Jakarta.* books.google.com.

Hanim, Wasifah. 2016. *Ekonomi Pembangunan.*

Hasanah, Herliyani, Anisatul Farida, and Pineda Prima Yoga. 2022. “Implementation of Simple Linear Regression for Predicting of Students’ Academic Performance in Mathematics.” *Jurnal Pendidikan Matematika (Kudus)* 5(1):38. doi: 10.21043/jpmk.v5i1.14430.

Hasibuan, Sanna Paija, Shalati Febjislami, and Irfan Suliansyah. 2022. “PENGARUH PUPUK KANDANG AYAM TERHADAP PERTUMBUHAN DAN KUALITAS BIJI TANAMAN SORGUM (Sorghum Bicolor L.).” *Jurnal Pertanian Presisi (Journal of Precision Agriculture)* 6(1):15–27. doi: 10.35760/jpp.2022.v6i1.5131.

Karami, Keyvan, Mahboubeh Akbari, Mohammad Taher Moradi, Bijan Soleymani, and Hossein Fallahi. 2021. “Survival Prognostic Factors in Patients with Acute Myeloid Leukemia Using Machine Learning Techniques.” *PLoS ONE* 16(7 July).

Lailatul Nikmah, Tiara, Risma Moulidya Syafei, Rini Muzayanah, Asharinnisa Salsabila, and Alya Aulia Nurdin. 2022. “Prediction of Used Car Prices Using K-Nearest Neighbour, Random Forest, and Adaptive Boosting Algorithm.” *International Conference on Optimization and Computer Application* 1(1 SE-Articles):17–22.

Putra, Indra Permana, I. Ketut, Gede Suhartana, and Bukit Jimbaran. 2022. “Perbandingan Akurasi Algoritma Regresi Linier, Regresi Polinomial, Dan Support Vector Regression Pada Model Sistem Prediksi Harga Rumah.” *Jnatia* 1(1):147–54.

Ruvananda, Adam Rahmat, and M. Taufiq. 2022. “Analisis Faktor-Faktor Yang Mempengaruhi Impor Beras Di Indonesia.” *Kinerja* 19(2):195–204. doi: 10.30872/jkin.v19i2.10924.

Saadah, Siti, and Haifa Salsabila. 2021. “Prediksi Harga Bitcoin Menggunakan Metode Random Forest.” *Jurnal Komputer Terapan* 7(1):24–32. doi: 10.35143/jkt.v7i1.4618.

Sedaghat, Azadeh, Mahmoud Shabanpour Shahrestani, Ali Akbar Noroozi, Alireza Fallah Nosratabad, and Hossein Bayat. 2022. “Developing Pedotransfer Functions Using Sentinel-2 Satellite Spectral Indices and Machine Learning for Estimating the Surface Soil Moisture.” *Journal of Hydrology* 606. doi: 10.1016/j.jhydrol.2021.127423.

Shahini, F., and N. Grgurić. 2021. “Prediction of Tool Wear after Machining.” *Ri-STEM-2021*.

Supriyanto, Sarjana, M. Ilhamsyah, and Ultach Enri. 2022. “Prediksi Harga Minyak Kelapa Sawit MenggunakanLinear Regression Dan Random Forest.” *Jurnal Ilmiah Wahana Pendidikan* 8(7):1–8. doi: 10.5281/zenodo.6559603.